KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Order batching in multi-server pick-and-sort warehouses

Inneke Van Nieuwenhuyse, René B.M. de Koster and Jan Colpaert

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0731

Order Batching in Multi-Server Pick-and-Sort Warehouses

Inneke Van Nieuwenhuyse

Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven

Naamsestraat 69, 3000 Leuven, Belgium

Inneke.vannieuwenhuyse@econ.kuleuven.be

René B.M. de Koster

RSM Erasmus University

PO Box 1738, 3000 DR Rotterdam, Netherlands

rkoster@rsm.nl

Jan Colpaert

Centre for Modeling and Simulation, European University College Brussels

Stormstraat 2, 1000 Brussels, Belgium

jan.colpaert@ehsal.be

ABSTRACT

In many warehouses, customer orders are batched to profit from a reduction in the order picking effort. This reduction has to be offset against an increase in sorting effort. This paper studies the impact of the order batching policy on average customer order throughput time, in warehouses where the picking and sorting functions are executed separately by either a single operator or multiple parallel operators. We present a throughput time estimation model based on Whitt's queuing network approach, assuming that the number of order lines per customer order follows a discrete probability distribution and that the warehouse uses a random storage strategy. We show that the model is adequate in approximating the optimal pick batch size, minimizing average customer order throughput time. Next, we use the model to explore the different factors influencing optimal batch size, the optimal allocation of workers to picking and sorting, and the impact of different order picking strategies such as sort-while-pick (SWP) versus pick-and-sort (PAS).

Keywords: Queuing, warehousing, order batching, order picking and sorting

1. INTRODUCTION

Warehouses play an important role in companies' supply chains. Among the many activities carried out in a warehouse, order picking - *the process of retrieving products from storage in response to a specific customer request* - is the most critical one: it has to be carried out in a short available time, meeting truck departure due times. Order picking may consume as much as 60% of all labor activities in the warehouse, and, for a typical warehouse, the cost of order picking is estimated to be as much as 55% of the total warehouse operating expense (Tompkins et al., 2003). For these reasons, warehousing professionals consider order picking as the highest-priority activity for productivity improvements.

Four operational decision problems influencing the performance of (manual) order picking systems have received attention from researchers (De Koster et al., 2006):

- *Storage assignment.* Storage assignment methods assign stock keeping units (SKUs) to storage locations. This assignment impacts the order-picking throughput time. The main storage policies mentioned in the literature are randomized, class-based and dedicated storage. The easiest storage method is to randomly allocate incoming products to available storage locations. However, we can reduce the expected travel time of a picking tour by locating high-demand products near the input/ output (I/O) point (or depot) of the warehouse, which can be done on a group or on item basis. In practice, pick-frequency class-based storage strategies (Hausman et al., 1976) are most popular. Such a strategy divides products and locations into classes, ranks product classes in decreasing order of pick frequency, and then assigns them in that order to the location classes nearest to the I/O point. A dedicated storage strategy (Caron et al., 1998, 2000) ranks the items individually to some criterion (for example pick frequency) and then assigns them in that order to the locations nearest to the I/O point. The cube-per-order index (COI) rule, which is attributed to Heskett (1964), is an example of such a dedicated storage strategy. The COI is the ratio of the space requirement (cube) of a SKU to its turnover rate.

- *Layout problem*. This is the problem of finding a good aisle configuration (i.e. the optimal number and length of aisles) minimizing order picking throughput time. Little research has been done in this area. Roodbergen (2001) proposes a non-linear objective function (i.e. average travel time in terms of number of picks per route and pick aisles) for determining the aisle configuration for random storage warehouses (including single and multiple blocks) that minimizes the average tour length. Also considering minimization of the average tour length

as the major objective, Caron et al. (2000) consider 2-block warehouses (i.e., one middle cross aisle) under the COI-based storage assignment. For small (up to 2-block) class-based storage warehouses, Le-Duc and De Koster (2005a) propose a travel time model and a local search procedure for determining optimal storage zone boundaries as well as the number of storage aisles.

- *Routing order pickers.* This well-researched problem considers the determination of the optimal sequence of visits to pick up a number of requested items as quickly as possible. A polynomial time optimal routing method for a single-block rectangular warehouse is due to Ratliff and Rosenthal (1983), and has been further extended to various layouts and working methods by several authors (Goetschalckx and Ratliff, 1988; De Koster and Van der Poort, 1998; Roodbergen and De Koster, 2001b). The disadvantage of exact algorithms is that they depend on the layout and depot location, and that the resulting routes may be too complicated for pickers to follow (Dekker et al., 2004). For large and more complicated layouts (more than two blocks) several heuristics are documented. The best routing heuristic known so far is probably the combined heuristic (Roodbergen and De Koster, 2001a). This method combines two basic methods: either traversing a visited aisle from one end to the other or entering and leaving the aisle from the same aisle's end. The choices are made by using dynamic programming.
- *Batching* and *zoning.* Batching determines which orders are released together. With batch picking, multiple orders are picked together in one pick tour and need to be sorted by order later. By sharing a pick tour, the average travel time per order is reduced. Basically, two criteria for batching exist: proximity of pick locations batching and time-window batching. Proximity batching refers to the clustering of a given number of orders based on retrieval locations (Hwang et al.,1988; Gibson and Sharp, 1992; Elsayed et al., 1993; Rosenwein, 1994; Elsayed and Lee 1996; De Koster et al., 1999; Gademann et al., 2001; Gademann and van de Velde, 2005). Time-window batching studies the order batching problem in a stochastic context. The number of orders per batch can be fixed or variable. Variable time-window batching groups all orders that arrive during the same time interval or window. With fixed-number-of-orders time-window batching, the time window is the variable length until a batch has a predetermined number of orders (Le-Duc and De Koster, 2007). Zoning is closely related to batching; it divides the pick area into sub-divisions (or zones), each with one or few pickers dedicated to it. The major advantages of zoning are: reduction of the travel time (because of the smaller traversed area and also the familiarity of the picker with the zone) and of the traffic congestion. Depending on the pick process sequence, zoning can be further

classified as progressive zoning or synchronized zoning. With progressive zoning, orders are sequentially picked zone by zone (this system is also called 'pick-and-pass'); a batch is finished when all (order) lines of the orders in the batch are picked. In contrast, in synchronized (or parallel) zoning, pickers in all zones can work on the same batch at the same time (Choe and Sharp, 1991). In synchronized zoning, the picking process must be followed by a sorting (and often also a packing) process, to group the items of the same order picked by the multiple pickers (Le-Duc and De Koster, 2005b). Zoning has received little attention in the literature despite its important impact on the performance of the order picking system. Choe et al. (1993) study the effects of three order picking strategies in an aisle-based system: single-order-pick, sort-while-pick, and pick-and-sort. They propose analytical tools for the planner to quickly evaluate various alternatives without using simulation.

Time-window batching is becoming more and more the rule in many warehouses, implying that orders arrive online and have to be processed as they arrive. This is partly due to an increased pressure on short delivery lead times: although customers can enter orders late, they still need to be picked, packed, and shipped the same day. Minimizing order throughput times is therefore an important objective in many warehouses. Order batching helps in achieving this objective. Several trade-offs exist in the order batching process: if batch sizes increase, both order picker travel time and batch start-up time per order decrease, but orders have to wait longer in the queue for batch completion. Also, larger batches imply longer processing times in the sorting and packing process, as all picked order lines have to be grouped and packed per order. This implies that pick batches must wait longer in the queue for sorting and packing.

In view of these trade-offs, we consider the question of finding an optimal batch size for the picking and sorting-and-packing process that minimizes the average order throughput time. Although the problem of finding good batch sizes incorporating picking, sorting and packing is quite common in warehouses, it is not in the literature. Only Choe et al. (1993), Chew and Tang (1999), and Le-Duc and De Koster (2007) address order batching in a stochastic (online) context. To our knowledge, this paper is the first to analytically study the impact of the consecutive sorting and packing processes on order batching decisions, for multi-server systems with general interarrival and service time distributions. We consider a setting with fixed number-of-orders batching, where the objective is to find that particular pick batch size (expressed in number of customer orders) that minimizes the average customer order throughput time. While actual pick batch sizes in practice are likely to vary from one pick batch to the next, the optimal pick batch size as obtained from the model can be considered as a target batch size to be used in real-life operations. We subsequently use the model to study a number of managerially relevant issues, such as which factors have a crucial impact on

the optimal pick batch size, how the workforce should be optimally allocated to picking and sorting operations, and whether or not a sort-while-pick policy may outperform a pick-and-sort policy.

The analytical model provides a decision support tool for management, enabling to compare different system alternatives without having to resort to lengthy simulations. It can be extended relatively easily to include other batching rules, such as batch sizing based on order lines rather than on number of customer orders. Our method is based on two-moment approximations for multi-server queues with batching and general service and arrival distributions and builds on the work of Le-Duc and De Koster (2007), who provide estimates for the first two moments of picking time for order batches. In the next section, we describe the order picking operation and introduce notations and assumptions. In section 3, we present the model for estimating the customer order throughput time as a function of batch size. Section 4 discusses the trade-offs that imply the existence of an optimal pick batch size, followed in section 5 by a validation of the model's accuracy in determining the optimal system behavior. Section 6 presents insights on the effect of different factors influencing optimal batch size, a comparison of various order picking policies like pick-and-sort versus sort-while-pick, and an evaluation of allocation of workers to picking and sorting stations. Section 7 summarizes the conclusions.

2. NOTATION AND ASSUMPTIONS

We consider a setting in which orders arrive online and are then batched for picking. Batched orders are picked in one pick tour in a rectangular 2-block warehouse, as sketched in Figure 1.

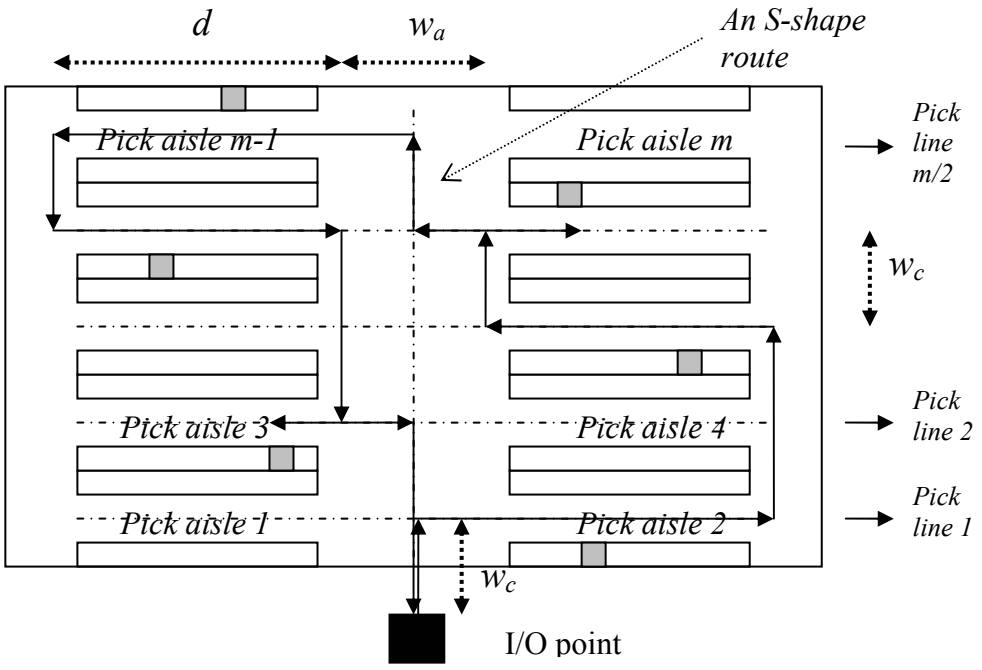


Figure 1: Detailed flowchart of a 2-block warehouse with an S-shape picking route

This layout is rather common in practice and used by Caron et al. (1998), among others. When an order batch is started for picking, setup time is required. This time includes the time for the picker to obtain a pick list, an empty pick cart and, at the end of the pick process, to drop-off the pick cart at an I/O point. Pickers walk along the aisles according to a certain routing policy (we assume S-shape routing) and can pick in the rack at the left and right side of the aisle without additional travel (this is called 'two-sided picking'). All rack levels can be reached by the picker without additional vertical travel (like in conventional shelf racks, or pallet racks where picking occurs from the bottom levels only).

As the picked items on the pick cart belong to multiple orders, they have to be sorted and, in most cases, packed by order. If the picker has already grouped the order lines on the pick cart per order, only consecutive packing is needed. Such a picking process is called 'sort-while-pick' (SWP) and is possible only if the number of orders in the batch is not too large, since sorted orders occupy more space, and to avoid sorting errors by the picker. In a pick-and-sort policy (PAS), the picker does not sort the orders during the picking process; sorting and packing happens at a second process. Setup times are required to retrieve the cart from the depot, and for retrieving packing instructions and materials.

We assume that the interarrival time of customer orders is generally distributed (only first and second moments are known). The number of order lines in a customer order follows a given discrete distribution with a minimum of one (e.g., shifted Poisson, shifted negative binomial). A fixed number of customer orders k is grouped into an order picking batch; a picker then collects the list with the order lines at the input point of the picking zone. Pickers travel the picking area according to the S-shape heuristic. The picking time per order line is random (denoted by X_p) with a general probability distribution (only first and second moments are known). Pickers travel at constant speed, and items in the warehouse are stored according to a random storage strategy. Pick batches that are complete are dropped off by the picker at the output point and moved to the sorting and packing area. The sorting plus packing time per order line is random (denoted by X_s) with a general probability distribution (only first and second moments are known). The picking and sorting processes are carried out by either single or multiple parallel operators. A setup time may be required both at picking and sorting (denoted by SU_p and SU_s, respectively). This setup time is assumed to be independent of the order batch size, and to be generally distributed (only first and second moments are known).

In the next section, we develop a queueing model to estimate the average customer order throughput time in terms of the order batch size. The following notation will be used:

d length (in travel time units) of a pick aisle

w_a width (in travel time units) of the cross aisle

w_c center-to-center distance (in travel time units) between two adjacent pick aisles

k total number of customer orders to be picked on a tour

Q total number of order lines to be picked on a tour (pick batch size or order batch size)

$P^k(n)$ probability that the pick batch size Q equals n order lines, given that it consists of k customer orders

m total number of pick aisles (integer and even)

p_i probability that a random order line item in the order batch is picked from pick aisle i ($i=1,\ldots,m$); $p_i = 1/m$ for the random storage strategy

Y interarrival time of customer orders

O customer order size

TR_{WA} travel time caused by traversing the pick aisles (within-aisle travel time)

TR_{CA} travel time caused by traversing the cross aisle (cross-aisle travel time)

SU_p setup time per pick batch at picking station

SU_s setup time per pick batch at sorting station

X_p picking time per order line

X_s sorting plus packing time per order line

n_p number of order pickers in the picking area

n_s number of sorting servers in the sorting area

ρ_p utilization rate of the picking process

ρ_s utilization rate of the sorting process

The notation $E(Z)$ denotes the expected value of random variable Z; $Var(Z)$ denotes its variance, $E(Z^2)$ its second moment, and c_Z^2 its squared coefficient of variation (SCV).

3. ESTIMATION OF AVERAGE CUSTOMER ORDER THROUGHPUT TIME

The average throughput time of a customer order ($E(W)$) can be written as a sum of different components:

- the average time that the order spends waiting for the pick batch to be formed (the average collection time), denoted by $E(W_{coll})$;
- the average time that the pick batch spends in queue at the picking area and the sorting area, denoted by $E(W_{q,p})$ and $E(W_{q,s})$ respectively;
- the average total service time needed for a pick batch in the picking area, $E(S_p)$, which consists of the average setup time $E(SU_p)$, the average travel time $E(TR)$ and the average picking time $E(W_p)$;

- the average total service time needed for a pick batch in the sorting area, $E(S_s)$, which consists of the average setup time $E(SU_s)$ and the average sorting time $E(W_s)$.

Each of these components (except the setup times) depends on the number of customer orders k that are collected into the batch. Hence, we may write the following expression:

$$E(W|k) = \underbrace{E(W_{coll}|k)}_{collection} + \underbrace{\underbrace{E(W_{q,p}|k)}_{waiting} + \underbrace{E(SU_p) + E(TR|k) + E(W_p|k)}_{service:E(S_p|k)}}_{picking\ area} + \underbrace{\underbrace{E(W_{q,s}|k)}_{waiting} + \underbrace{E(SU_s) + E(W_s|k)}_{service:E(S_s|k)}}_{sorting\ area} \quad (1)$$

For a given number of customer orders k in the order batch, the average collection time $E(W_{coll}|k)$ of an arbitrary order in the batch can be written in a straightforward manner:

$$E(W_{coll}|k) = \frac{k-1}{2}E(Y) \quad (2)$$

In what follows, we first discuss in detail the first and second moments of the service time for a pick batch in the picking and sorting area; next, we present the queueing expressions which are used to approximate $E(W_{q,p})$ and $E(W_{q,s})$.

3.1 Average and Variance of Service Time in the Sorting Area

The service time for a pick batch in the sorting area, S_s, consists of the setup time and the sorting time:

$$S_s = SU_s + W_s \quad (3)$$

As SU_s and W_s are independent, the average and variance of S_s can be written conditional upon the number of customer orders k contained in the pick batch:

$$\begin{aligned} E(S_s|k) &= E(SU_s) + E(W_s|k) \\ Var(S_s|k) &= Var(SU_s) + Var(W_s|k) \end{aligned} \quad (4)$$

The average and variance of the setup time SU_s are given, and are independent of the pick batch size Q. As the sorting times for individual order lines in the order batch (X_s) are IID distributed, the sorting time W_s for the order batch can be interpreted as the sum of a random number (Q) of IID random variables (X_s), with Q and X_s independent. For a specific number of customer orders k, the first moment of W_s is then given by:

$$E(W_s|k) = E(Q|k)E(X_s) \quad (5)$$

where $E(Q|k)$ refers to the expected number of order lines in the pick batch, given that it consists of k customer orders. It is given by:

$$E(Q|k) = \sum_{n=k}^{\infty} nP^k(n) = kE(O) \quad (6)$$

where $P^k(n)$ denotes the probability that the order batch size Q, for a given number of customer orders k, consists of n order lines. The expression for $P^k(n)$ is given by the k-fold convolution of the probability mass function of the customer order size O. Note that the number of order lines in the pick batch will be larger than or equal to k, as a single customer order consists of at least one order line.

The variance of W_s, for a given number of customer orders k in the pick batch, can be written as (Blumenfeld, 2001):

$$Var(W_s|k) = E(Q|k)Var(X_s) + Var(Q|k)\left[E(X_s)\right]^2 \tag{7}$$

where $Var(Q|k)$ is given by:

$$Var(Q|k) = \sum_{n=k}^{\infty} n^2 P^k(n) - \left[E(Q|k)\right]^2 \tag{8}$$

By means of expressions (4) to (8), the average and variance of S_s in expression (3) can be determined for any arbitrary number of customer orders k contained in the pick batch, and any arbitrary discrete distribution of the customer order size.

3.2 Average and Variance of Service Time in the Picking Area

The service time for a pick batch in the picking area, S_p, is given by the sum of the setup time, the actual picking time, and the travel time of the order picker:

$$S_p = SU_p + W_p + TR \tag{9}$$

As the three random variables are mutually independent, the average and variance of S_p are given by:

$$\begin{aligned} E(S_p|k) &= E(SU_p) + E(W_p|k) + E(TR|k) \\ Var(S_p|k) &= Var(SU_p) + Var(W_p|k) + Var(TR|k) \end{aligned} \tag{10}$$

for any given number of customer orders k contained in the pick batch. The average and variance of SU_p are given, and are independent of the pick batch size Q. The random variable W_p depends on Q in a way that is similar to W_s (as discussed in section 3.1). Hence, for a given number of customer orders k in the pick batch, the average and variance of W_p are given by:

$$\begin{aligned} E(W_p|k) &= E(Q|k)E(X_p) \\ Var(W_p|k) &= E(Q|k)Var(X_p) + Var(Q|k)\left[E(X_p)\right]^2 \end{aligned} \tag{11}$$

with $E(Q|k)$ given by expression (6), and $Var(Q|k)$ by expression (8). The average and variance of the travel time TR depend not only on the pick batch size, but also on three additional factors: warehouse layout, storage strategy, and routing policy of the order pickers through the warehouse. For a 2-block rectangular warehouse with S-shape routing policy, as we are considering here, an

approximation for the first and second moment of travel time has been described in Le-Duc and De Koster (2007). The resulting expressions depend on the number of order lines *n* contained in the pick batch, and are valid for both the random storage policy and an ABC-storage policy without partial-aisle assignment. Appendix A gives an overview of their results. In what follows, we present simplified versions of these expressions for the case of a random storage strategy and adapt them to incorporate the impact of variable pick batch sizes.

As shown in the appendix (see expression (A.2)), the first moment of travel time for a given pick batch size of *n* order lines consists of four components: $E[TR|n] = E[TR_{WA}|n] + E[TR_{CA}|n] + E[AT_1|n] + E[AT_2|n]$. Both $E[AT_1|n]$ and $E[AT_2|n]$ are adjustment terms, given by expressions (A.3) and (A.4) in the appendix. The expressions for the average within-aisle travel time $E[TR_{WA}|n]$ and the average cross-aisle travel time $E[TR_{CA}|n]$ given in (A.1) can be further simplified in case of a random storage strategy (as $p_i = 1/m$). For a given pick batch size of *n* order lines, we obtain:

$$\begin{aligned} E[TR_{WA}|n] &= dm\left[1-\left(1-\frac{1}{m}\right)^n\right] \\ E[TR_{CA}|n] &= 2w_c\left[\frac{m}{2}-\sum_{l=1}^{m/2-1}\left(\frac{(2m-1)l}{m^2}\right)^n\right] = 2w_c\left[\frac{m}{2}-\frac{(2m-1)^n}{m^{2n}}\sum_{l=1}^{m/2-1}l^n\right] \end{aligned} \tag{12}$$

The first moment of travel time for a pick batch consisting of *k* customer orders is then given by:

$$E[TR|k] = \sum_{n=k}^{\infty} P^k(n) E[TR|n] \tag{13}$$

where $P^k(n)$ again denotes the probability that the pick batch size consists of *n* order lines, given that *k* customer orders are grouped in the order batch. The second moment of travel time for a given pick batch size of *n* order lines is given by (see expression (A.6) in appendix):

$$E[TR^2|n] = d^2 E[J^2|n] + (2w_c)^2 E[L^2|n] + 4w_c d E[JL|n] \tag{14}$$

where *J* refers to the number of aisles visited, and *L* to the pick line (see Figure 1) of the farthest visited aisle. In case of a random storage policy, the expressions in (A.7) reduce to:

$$\begin{aligned} E[J^2|n] &= m\left((m-1)\left(\frac{m-2}{m}\right)^n + (1-2m)\left(\frac{m-1}{m}\right)^n + m\right) \\ E[L^2|n] &= (m/2)^2 - \sum_{i=1}^{m/2-1}(2i+1)\left(\frac{i(2m-1)}{m^2}\right)^n = (m/2)^2 - \left(\frac{2m-1}{m^2}\right)^n \sum_{i=1}^{m/2-1}(2i+1)i^n \\ E[JL|n] &= 2^{n+1}\sum_{l=1}^{m/2} l\left[\left(1-\left(1-\frac{1}{2l}\right)^n\right)\frac{l^{n+1}}{m^n} + \left(1-\left(1-\frac{1}{2(l-1)}\right)^n\right)\frac{(l-1)^{n+1}}{m^n}\right] \end{aligned} \tag{15}$$

The variance of travel time, for a given number of order lines *n*, can be approximated using expressions (12) and (14):

$$Var\left[TR|n\right] = E\left[TR^2|n\right] - \left(E\left[TR_{CA}|n\right] + E\left[TR_{WA}|n\right]\right)^2 \tag{16}$$

The variance of travel time for a pick batch consisting of *k* customer orders is then given by:

$$Var\left[TR|k\right] = \sum_{n=k}^{\infty} P^k(n) E\left[TR^2|n\right] - \left[\sum_{n=k}^{\infty} P^k(n)\left(E\left[TR_{CA}|n\right] + E\left[TR_{WA}|n\right]\right)\right]^2 \tag{17}$$

Using expressions (11), (13) and (17) in expression (10), we have obtained an approximation for $E\left[S_p|k\right]$ and $Var\left[S_p|k\right]$.

3.3 Approximation for $E(W_{Q,p})$ and $E(W_{Q,s})$

In the system under study, pick batches may have to wait in queue both at the picking area and at the sorting area. If we regard picking and sorting as separate stations consisting of either a single resource or multiple parallel resources, we may use single-class *G*/*G*/1 and *G*/*G*/*m* queueing expressions to approximate $E(W_{q,p})$ and $E(W_{q,s})$.

3.3.1 Single server stations

In case of a single-server station, we rely upon the approximation of Whitt (1983) (see Appendix B). Assuming that *k* customer orders are contained in a pick batch, we can write the following for $E(W_{q,p})$:

$$E(W_{q,p}|k)_{G/G/1} = \begin{cases} \dfrac{\rho_p^2\left(ca_p^2 + cs_p^2\right)}{2\lambda_p\left(1-\rho_p\right)} & \text{if } \mathrm{ca}_\mathrm{p}^2 \geq 1 \\ \dfrac{\rho_p^2\left(ca_p^2 + cs_p^2\right)}{2\lambda_p\left(1-\rho_p\right)} \exp\left\{\dfrac{-2\left(1-\rho_p\right)\left(1-ca_p^2\right)^2}{3\rho_p\left(ca_p^2 + cs_p^2\right)}\right\} & \text{otherwise} \end{cases} \tag{18}$$

where λ_p denotes the arrival rate of pick batches at the picking area, ca_p^2 denotes the SCV of interarrival times of pick batches at the picking area, cs_p^2 refers to the SCV of pick batch service time and ρ_p to the utilization of the picking server. These parameters depend on the number of customer orders *k* being grouped in the pick batch:

$$\lambda_p = \frac{1}{kE(Y)}, ca_p^2 = \frac{kVar(Y)}{\left(kE(Y)\right)^2}, cs_p^2 = \frac{Var(S_p|k)}{\left(E(S_p|k)\right)^2}, \rho_p = \lambda_p E(S_p|k) \tag{19}$$

with $E(S_p|k)$ and $Var(S_p|k)$ given by expression (10). Note that the expression for ca_p^2 assumes that customer order interarrival times *Y* are IID distributed. The average waiting time of a pick batch at the sorting server can be derived in an analogous way:

$$E(W_{q,s}|k)_{G/G/1} = \begin{cases} \dfrac{\rho_s\left(ca_s^2 + cs_s^2\right)}{2\lambda_s\left(1-\rho_s\right)} & \text{if } ca_s^2 \geq 1 \\ \dfrac{\rho_s^2\left(ca_s^2 + cs_s^2\right)}{2\lambda_s\left(1-\rho_s\right)} \exp\left\{\dfrac{-2\left(1-\rho_s\right)\left(1-ca_s^2\right)^2}{3\rho_s\left(ca_s^2 + cs_s^2\right)}\right\} & \text{otherwise} \end{cases} \quad (20)$$

with

$$\lambda_s = \frac{1}{kE(Y)}, cs_s^2 = \frac{Var(S_s|k)}{\left(E(S_s|k)\right)^2}, \rho_s = \lambda_s E(S_s|k) \quad (21)$$

Here, $E(S_s|k)$ and $Var(S_s|k)$ are given by expression (4) above. The expression for λ_s follows from the conservation of flow in the system: this implies λ_s is equal to λ_p. As the sorting operation takes place after the picking operation, the SCV of interarrival times of pick batches at the sorting area (ca_s^2) obviously equals the SCV of interdeparture times of pick batches from the picking area (which we will denote by cd_p^2). When the picking station is a single server station, ca_s^2 can be approximated by the following linking equation (Marshall, 1968):

$$ca_s^2 = cd_p^2 = ca_p^2 + 2\rho_p^2 cs_p^2 - 2\rho_p(1-\rho_p)\frac{E(W_{q,p}|k)}{E(S_p|k)} \quad (22)$$

3.3.2 Multiserver stations

In case of a multi-server station, we approximate $E(W_q)$ by the expression developed by Whitt (1993), which is given in Appendix C. Given the number of servers n_s and n_p, and the parameters derived in (18) and (20) above, the expressions in Appendix C can be directly applied, yielding approximations for $E(W_{q,p})_{G/G/n_p}$ and $E(W_{q,s})_{G/G/n_s}$. The only additional expression needed is an approximation for the interdeparture time of pick batches from the multiserver picking area, as this will be equal to ca_s^2 due to the linking property. When the picking station is a multiserver station, ca_s^2 can be approximated by the following linking equation (Whitt, 1983; Hopp and Spearman, 2000):

$$ca_s^2 = cd_p^2 = 1 + (1-\rho_p^2)(ca_p^2 - 1) + (\rho_p^2/\sqrt{n_p})(cs_p^2 - 1) \quad (23)$$

4. IMPACT OF PICK BATCH SIZE ON AVERAGE ORDER THROUGHPUT TIME: ILLUSTRATION AND DISCUSSION

In this section, we apply the model to three warehouse instances, based on the examples used in Chew and Tang (1999) and Le-Duc and de Koster (2007). The parameters are listed in Table 1. It is assumed that the customer order size follows a shifted Poisson distribution, and that the picking station consists of 2 parallel servers.

The model is applied for pick batch sizes ranging from k=2 to k=10 customer orders[1]. In Figure 2, the resulting average throughput time of a customer order $E(W)$ is compared to discrete-event simulation results, obtained from runs consisting of 1000 pick batch sizes for every particular setting. The figure reveals that $E(W)$ is convex in k, implying that there exists an optimal number of customer orders per pick batch k_{opt}, which yields a minimum average throughput time.

Parameter	Level	Parameter	Level
m	6,10,16 aisles	w_a	6 seconds
Y	EXPO(150) seconds	w_c	10 seconds
O	1+Poisson(2)	SU_p	180 seconds
n_p	2	SU_s	0 seconds
n_s	1	X_p	EXPO(12) seconds
d	30 seconds	X_s	EXPO(10) seconds

Table 1: Parameters for the simulation setting

Figure 2: Average throughput time $E(W)$ for a customer order (in minutes), in terms of the number of customer orders (k) per pick batch

[1] It turns out that a pick batch size of k=1 customer order is infeasible in this setting (at k=1, $\rho_p \geq 1$ yielding an unstable system).

This confirms earlier results for the single-server settings with deterministic customer order sizes (Chew and Tang, 1999 and Le-Duc and de Koster, 2007). The convex relationship is the result of a trade-off between two underlying effects: the saturation effect and the batching effect (Karmarkar et al., 1985; Karmarkar et al., 1985; Karmarkar, 1987). To understand these effects, we refer to expression (1), which decomposes the average throughput time of a customer order in three elements: the collection time, the waiting time for service, and the service time. When the batch size is small, the collection time and service time are small, but the waiting time will be large. Indeed, small pick batch sizes require frequent setups, which cause the utilization of the system to increase, causing congestion: the pick batches will need to wait longer in queue in front of the picking and sorting process. The system becomes saturated (hence the term saturation effect). Note that the congestion effect in our model refers to the time that pick batches wait in queue in front of both processes; it does not refer to traffic congestion that might occur due to an excessive number of pickers being present in a given aisle. Traffic congestion is ignored in the current model, as we assume that pickers travel at constant speed (see Section 2).

In contrast, when the pick batch size is large, the waiting time will be small, but the collection time and service time are large. This is referred to as the batching effect. At high values of k, $E(W)$ increases almost linearly in k. This is intuitive, as from a certain value of k the probability that the pickers will need to travel the entire warehouse approaches one, such that TR becomes independent of k. In these cases, the impact of k on $E(S_p)$ and $E(S_s)$ will be largely due to the average picking time $E(W_p)$ and the average sorting time $E(W_s)$ respectively, which are approximately linear in k.

5. VALIDATION

In this section, we validate the model developed in Section 3 by means of discrete event simulation, in two respects: (i) the precision in approximating the average customer order throughput time $E(W)$ for an arbitrary pick batch size; (ii) the precision in approximating the optimal pick batch size k_{opt} and the corresponding optimal average order throughput time $E(W)_{opt}$. We studied a fictitious warehouse, of which the technical parameters (i.e., distance parameters and setup and processing characteristics in the picking and sorting area) are given, and which is confronted with a given customer order pattern. The actual parameters are shown in Table 2. For this setting, the following factors were varied: the number of pickers and sorters (n_p=6 and n_s=3, n_p=4 and n_s=2 or n_p=2 and n_s=1), and the number of warehouse aisles m (m=4, 8, 12, or 20). The customer order size O follows a shifted Poisson distribution: $O \sim 1+\text{POISS}(b)$, where the parameter b is set equal to 0.5, 1, 2, or 3. The different combinations of these factors give rise to 48 simulation scenarios. For every scenario, k (the total number of customer orders to be picked on a tour) was varied from 1 to 15, yielding 720 potential simulation runs. For certain scenarios however, low values of k yield an infeasible setting

as it causes the pickers to be utilized beyond capacity ($\rho_p \geq 1$ according to expression (19)). This occurs in 94 out of 720 runs. Hence, only 626 runs are retained in the validation experiment.

Parameter	Value	Parameter	Value
Y	LOGN with average = 50 seconds, SCV=4	SU_p	GAMM with average= 60 seconds, SCV=2
w_a	6 seconds	SU_s	LOGN with average = 30 seconds, SCV=1
w_c	10 seconds	X_p	LOGN with average = 8 seconds, SCV=4
d	30 seconds	X_s	GAMM with average = 10 seconds, SCV=0.5

Table 2: Parameters for the validation experiment

A preceding analysis of simulation output based on 10 replications (by means of Welch's method, see Law and Kelton 2000), revealed that the system exhibits a very long transient phase (up to around 250000 customer orders) in highly utilized settings. Due to this long transient period, application of the replication/deletion approach is unpractical, as it would render the simulation effort prohibitive. To rule out transient effects and at the same time keep the simulation effort within feasible limits, we decided to make a single long run (700000 customer orders) for every setting, taking only the data relating to the second half of every simulation's runlength into account for validation.

For every setting, the average throughput time of a customer order as observed in the simulation ($E(W)_{simul}$) was compared to the average throughput time obtained in the queuing model ($E(W)_{analytic}$). The resulting relative error is determined as $\varepsilon = \frac{E(W)_{analytic} - E(W)_{simul}}{E(W)_{simul}}$.

Figure 3 shows a scatterplot of ε in terms of the utilization of the pickers. From this plot, we may draw several conclusions. Firstly, the queuing model in general tends to overestimate the average throughput time of customer orders. Secondly, while the performance of the queuing model is satisfactory at low to moderate levels of ρ_p (ε <0.15 as long as $\rho_p \leq 0.75$), it deteriorates as ρ_p increases.

This may have several causes. On the one hand, it may be due to the fact that the approximation for $E(W_q)$ in case of a *G/G/m* station (as described in Appendix C) tends to deteriorate at high utilization levels. On the other hand, it may be due to small errors on the approximation for average travel time in the picking area (expression (A.2) in Appendix A). The results from the validation experiment showed that the average travel time is in general slightly overestimated; the maximum

approximation error stays very low however at ca +2%. However, as the expression for the waiting time in queue for a *G/G/m* station has a factor $(1-\rho)$ in the denominator (see expressions (C.1) and (C.2) in Appendix C), such small errors in the travel time estimate may cause large deviations in the waiting time estimate, particularly at high utilization levels. Further analysis reveals that ε depends on the average number of picks per aisle (ppa): values for ε larger than 0.2 are only observed in settings with ppa≤1. This is intuitive, as the average travel time then constitutes a substantial part of the picker's utilization, increasing the risk for overestimations. From our observations, we may conclude that the queuing model is generally reliable in settings with low to moderate utilization, but that results should be interpreted with caution in highly utilized settings.

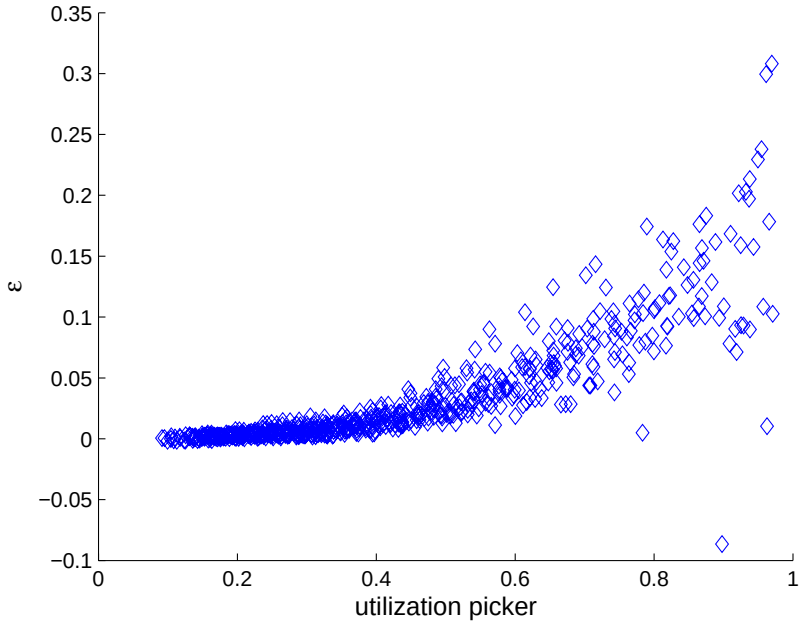


Figure 3: Relative error ε observed in the validation experiment, in terms of the pickers' utilization

Appendix D shows scatter plots of the actual values in minutes for the throughput time and its components as obtained from the simulation and the queueing model. The figures reveal that the errors in average throughput time are caused mainly by errors in the estimated waiting time in the system.

As the remainder of the paper focuses on analyzing the system's behavior with optimal pick batch sizing, we are particularly interested in the precision of the model in approximating the optimal pick batch size k_{opt} and the corresponding optimal average order throughput time $E(W)_{opt}$. The validation experiment showed that in 36 out of the 48 scenarios studied, the optimal pick batch size determined by the queuing model ($k_{opt,analytic}$) coincides with the optimum determined by simulation ($k_{opt,simul}$); the approximation error ε in these settings averages only +6.37% (varying between

1.12% and +16,17%). Table 3 lists the results for the 12 scenarios in which $k_{opt,analytic}$ differs from $k_{opt,simul}$. The table reveals that the simulated E(W) in the optimum predicted by the queueing model (column C) is only marginally higher than the simulated E(W) in $k_{opt,simul}$ (column B), indicating that the simulated throughput time curve is very flat near the optimum. Despite the fact that $k_{opt,analytic}$ does not coincide with $k_{opt,simul}$, the relative error ε on $E(W)_{opt}$ remains reasonable, averaging +9.32% (column D). Based on these observations, we consider the model to be adequate for our purposes, i.e. the study of the system's behavior at the optimum, as is done in Section 6.

m	$E(O)$	n_p	n_s	$k_{opt,analytic}$	$k_{opt,simul}$	$E(W)_{analytic}$ at $k_{opt,analytic}$ (minutes) *(A)*	$E(W)_{simul}$ at $k_{opt,simul}$ (minutes) *(B)*	$E(W)_{simul}$ at $k_{opt,analytic}$ (minutes) *(C)*	$\varepsilon = \frac{(B)-(A)}{(A)}$
4	2	2	1	5	4	11.18	10.12	10.13	+10.53%
4	4	6	3	2	1	7.28	7.02	7.15	+3.73%
8	1.5	2	1	6	5	12.87	11.52	11.78	+11.74%
8	2	2	1	7	6	15.82	14.35	14.50	+10.22%
8	4	2	1	9	8	32.66	29.68	30.54	+10.05
12	1.5	2	1	7	6	15.71	14.30	14.41	+9.9%
12	4	2	1	12	11	40.07	35.19	35.95	+13.89%
20	1.5	2	1	9	8	21.05	19.28	19.35	+9.20%
20	2	4	2	4	3	13.22	12.18	12.48	+8.54%
20	2	2	1	11	10	27.16	25.09	25.22	+8.24%
20	3	4	2	5	4	18.70	17.25	17.62	+8.41%
20	4	4	2	6	5	24.39	22.70	22.90	+7.45%

Table 3: Performance of the queuing model in the optimum, for the scenarios in which k_{opt} queue differs from k_{opt} simul

6. INSIGHTS

6.1. Factors Impacting the Optimal Pick Batch Size

In this section, we aim to investigate which factors impact k_{opt} and $E(W)_{opt}$. To also study the impact of the variability in customer order quantity, we determine the optimal behavior for the same scenarios as those described in Section 5, but now assuming that the customer order quantity O is deterministic: $E(O)=0$ and $Var(O)=0$. Note that, in a deterministic setting, customer order quantities should be integer: hence, we may only use the scenarios with $E(O)$=2, 3 or 4.

Table 4 compares the optimal k values for the scenarios in which O is deterministic ($k_{opt,DET}$ and $E(W)_{opt,DET}$) with the optimal k values obtained when O follows a shifted Poisson distribution ($k_{opt,POISS}$ and $E(W)_{opt,POISS}$). The table reveals that, for the settings studied, the optimal value of k was in general not impacted by the variability in the customer order quantity: $k_{opt,DET}$= $k_{opt,POISS}$,

except for the case m=4, $E(O)$=4, n_p=4 and n_s=2. The differences in the optimal average throughput times are negligibly small.

By contrast, the size of k (and the value of $E(W)_{opt}$) is heavily impacted by the number of aisles in the warehouse: a larger number of aisles induces a larger optimal value for k, particularly when the number of pickers is limited (n_p=2). This observation is likely linked to the assumption of a random storage strategy in the warehouse: indeed, a random storage strategy may imply large travel times in big warehouses, leading to inefficient use of the pickers' capacities (and, consequently, large utilizations) when the number of orders to be picked on a tour is small. Hence, as warehouses become larger, the travel times of the pickers exert an upward pressure on the optimal value of k. This observation also explains why the optimal k-values decrease as the number of pickers increases: indeed, enlarging the number of parallel pickers increases capacity. Though the average travel time on a single tour remains the same, the utilization of the pickers drops considerably, implying that smaller values of k (and hence, smaller pick batch sizes) become feasible and more attractive.

		n_p=2, n_s=1		**n_p=4, n_s=2**		**n_p=6, n_s=3**	
E(O)	***m***	$k_{opt,DET}$ $E(W)_{opt,DET}$	$k_{opt,POISS}$ $E(W)_{opt,POISS}$	$k_{opt,DET}$ $E(W)_{opt,DET}$	$k_{opt,POISS}$ $E(W)_{opt,POISS}$	$k_{opt,DET}$ $E(W)_{opt,DET}$	$k_{opt,POISS}$ $E(W)_{opt,POISS}$
2	4	5 11.19	5 11.19	2 6.25	2 6.12	1 4.36	1 4.21
	8	7 15.88	7 15.82	2 8.19	2 8.00	1 5.70	1 5.42
	12	8 19.82	8 19.73	3 9.90	3 9.80	2 7.18	2 7.10
	20	11 27.24	11 27.16	4 13.31	4 13.22	2 8.95	2 8.81
3	4	5 15.37	5 15.49	2 7.98	2 7.80	1 5.86	1 5.55
	8	8 21.55	8 21.62	3 11.05	3 10.90	2 7.89	2 7.79
	12	10 27.66	10 27.67	3 13.92	3 13.68	2 9.14	2 9.01
	20	14 39.28	14 39.25	5 18.81	5 18.70	2 12.52	2 12.24
4	4	7 26.60	7 27.09	3 9.80	2 9.66	2 7.34	2 7.27
	8	9 32.23	9 32.68	3 13.91	3 13.68	2 9.41	2 9.27
	12	12 39.71	12 40.07	4 17.45	4 17.29	2 11.48	2 11.27
	20	15 57.27	15 57.47	6 24.52	6 24.39	3 15.74	3 15.59

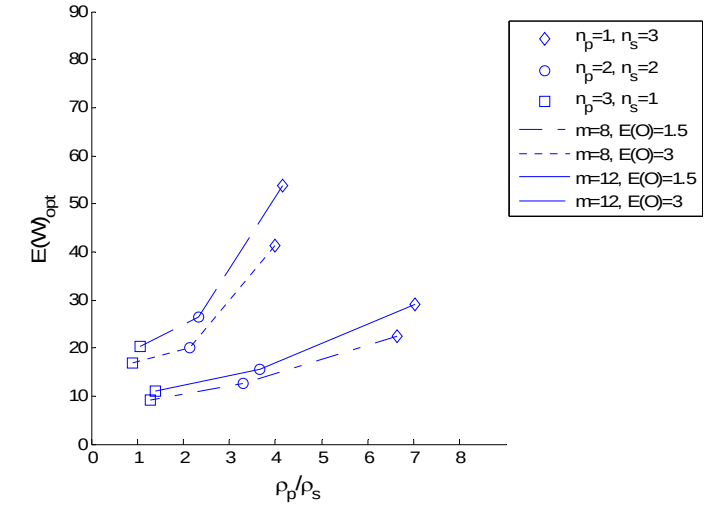
Table 4: Optimal number of orders k to be grouped in a pick batch, for both shifted Poisson distributed ($k_{opt,POISS}$) and deterministic ($k_{opt,DET}$) customer order quantities.

6.2. Optimal Allocation of Workforce to Picking and Sorting Operations

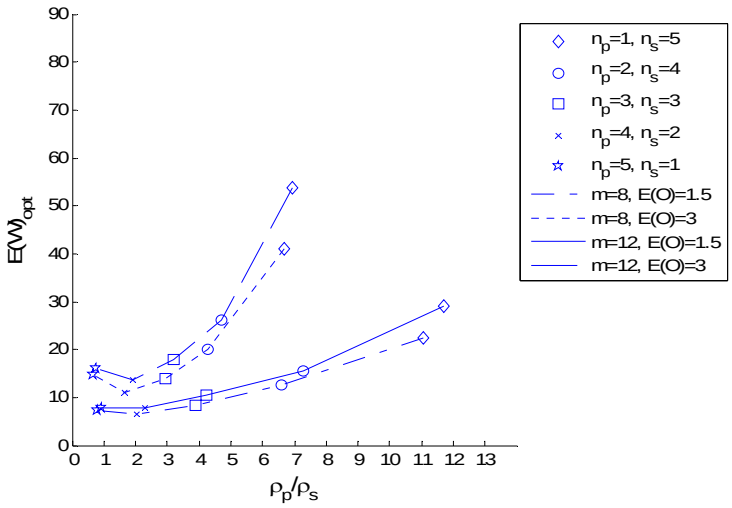
The queuing model developed in Section 3 can be used to investigate the allocation of a given workforce to the picking and sorting operations in order to minimize the average customer order throughput time. For this purpose, we studied a subset of the scenarios used in the validation experiment (see Section 5, Table 2 for the relevant input data). The number of aisles was restricted to *m=8* aisles or *m=12* aisles, and the customer order size *O* follows a shifted Poisson distribution: $O \sim 1+\text{POISS}(b)$, where the parameter *b* is either 0.5 or 2. An exhaustive experiment was performed for these settings, assuming a total workforce (n_{tot}) of four, six or eight people. For all possible combinations of n_s and n_p, the optimal pick batch size (and corresponding optimal customer order throughput time $E(W)_{opt}$) was determined, along with the utilization of the pickers (ρ_p) and of the sorters (ρ_s) in the optimum.

Figure 4 shows the resulting optimal customer order throughput times $E(W)_{opt}$ obtained for different allocations of n_s and n_p, for n_{tot}= 4 (a), n_{tot}=6 (b) and n_{tot}=8 (c). From the figure, we may derive that $E(W)_{opt}$ can be substantially reduced by allocating n_s and n_p in such a way that the picking and sorting operations become more closely balanced in the corresponding optimum (ρ_p / ρ_s close to one). Within any given scenario (i.e., for a given value of n_{tot}), the optimal allocation of n_s versus n_p turned out to be the same regardless of the number of aisles or the probability distribution of customer order size (for n_{tot} = 4 the optimal allocation is n_p=3 and n_s=1 in all cases, for n_{tot} = 6 the optimal allocation is n_p= 4 and n_s= 2, and for n_{tot} = 8 we obtain n_p= 6 and n_s= 2).

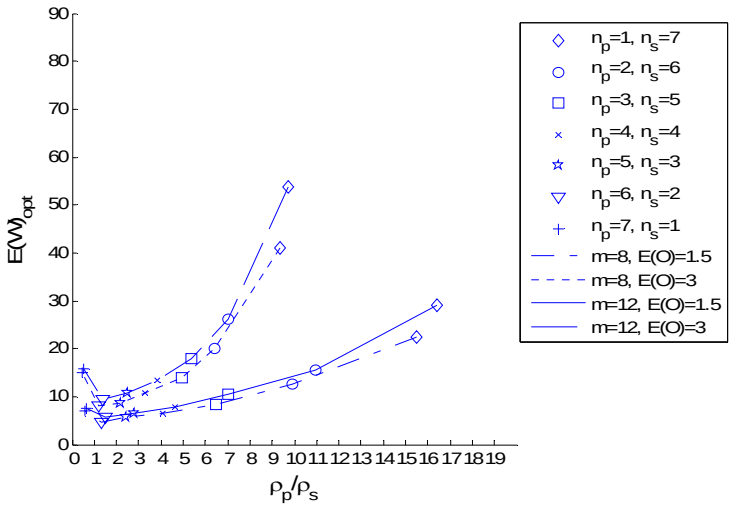
This result highlights the importance of the personnel allocation decision: while optimal batch sizing can be used to fine-tune performance for a given setting, it is unable to compensate for the inherent performance loss suffered from ill-considered allocation decisions. Moreover, the optimal personnel allocation appears to depend primarily on the total number of personnel available, and to be rather insensitive to changes in the warehouse's size or customer order patterns. Hence, personnel allocation should be considered as the first and most crucial optimization decision in a real-life setting; while this decision is robust in relation to changes in the environment, batch sizing decisions can be easily updated to changes in either customer order patterns or changes to the warehouse's size.



(a)



(b)



(c)

Figure 4: Optimal throughput time *E(W)* obtained for different allocations of n_s and n_p, for n_{tot}= 4 (a), n_{tot}=6 (b) and n_{tot}=8 (c)

6.3. Comparison of PAS and SWP Strategy

Finally, we used the model developed in section 3 to gain insight into the following question: given that we dispose of a given workforce n_{tot}, and that the pick batch size is determined optimally, what is the difference in performance between a PAS policy and a SWP policy for different personnel allocations?

As mentioned in section 2, the SWP policy implies that the picker sorts the different line items into customer orders before handing them over to the sorting operation; the sorting personnel primarily takes care of packaging the customer orders and preparing them for shipment. The impact of a move from PAS to SWP on the operating characteristics of the system (average and variability of SU_p, X_p, SU_s and X_s) is hard to evaluate in general, as it will depend on the actual system being studied. To compare the performance of PAS with SWP, we used the same scenarios as in section 6.2, assuming that $E(X_s)$ decreases with 30% (so $E(X_s)$ drops from 10 sec to 7 sec) while $E(X_p)$ increases accordingly from eight seconds to 11 seconds. Setup time characteristics and SCV's of both setup times and processing times were left unchanged. Table 5 compares the results for $E(W)_{opt}$ obtained with the PAS policy (as shown in Figure 4) and the SWP policy, along with the optimal value of k. From this table, we observe the following:

(i) For small values of n_p, the same optimal value of k and the same value of $E(W)_{opt}$ is obtained regardless of n_{tot}, for both PAS and SWP. This is an intuitive result: the understaffing of the picking operation implies that the picking operation is the bottleneck, and hence constrains the system. Adding personnel without changing n_p is futile in such a setting: it merely leads to an unnecessarily high number of sorters, without impacting system performance.

(ii) Moving from a PAS policy to a SWP policy tends to leave the optimal value of k for a given workforce allocation unchanged, except in the following cases:

- When n_p is small, a change towards a SWP policy increases the optimal value of k. This is again intuitive: when n_p is small, the picking operation is the bottleneck operation, so the increase in X_p due to the SWP policy only further aggravates the bottleneck situation, and hence further constrains the system.
- In case of abundant personnel (n_{tot}=6 or 8), with only one member of staff allocated to the sorting area (n_s=1), the SWP policy *may* allow a lower optimal value of k than the PAS policy. As the sorting operation is understaffed in these settings, the reduction in X_s caused by the SWP policy now alleviates the bottleneck operation and enables smaller pick batch sizes throughout the entire system.

(iii) For a given value of n_{tot} and a given workforce allocation, the SWP policy can only (marginally) outperform the PAS policy in terms of $E(W)_{opt}$ in those settings where the number of sorters is small (n_s= 1 or 2). When n_p is small, PAS consistently outperforms SWP.

			PAS			SWP		
m	***E(O)***	n_p	$n_{tot}=4$	$n_{tot}=6$	$n_{tot}=8$	$n_{tot}=4$	$n_{tot}=6$	$n_{tot}=8$
8	**1.5**	**1**	22.54 (15)	22.54 (15)	22.54 (15)	25.22 (18)	25.22 (18)	25.22 (18)
		2	12.60 (6)	12.57 (6)	12.57 (6)	13.29 (6)	13.27 (6)	13.27 (6)
		3	**9.14** (3)	8.39 (3)	8.38 (3)	**9.19** (3)	8.75 (3)	8.75 (3)
		4	/	**6.47** (2)	6.37 (2)	/	**6.61** (2)	6.55 (2)
		5	/	7.20 (2)	5.72 (2)	/	6.66 (2)	5.76 (2)
		6	/	/	**4.68** (1)	/	/	**4.67** (1)
		7	/	/	7.10 (2)	/	/	6.47 (2)
8	**3**	**1**	41.15 (22)	41.14 (22)	41.14 (22)	61.03 (33)	61.03 (33)	61.03 (33)
		2	20.09 (8)	19.95 (8)	19.95 (8)	22.38 (9)	22.32 (9)	22.32 (9)
		3	**16.89** (5)	14.02 (4)	14.00 (4)	**15.84** (5)	14.99 (5)	14.98 (5)
		4	/	**10.90** (3)	10.68 (3)	/	**11.35** (3)	11.25 (3)
		5	/	14.77 (3)	8.58 (2)	/	11.36 (2)	8.97 (2)
		6	/	/	**8.09** (2)	/	/	**8.03** (2)
		7	/	/	14.90 (3)	/	/	10.49 (2)
12	**1.5**	**1**	29.19 (20)	29.19 (20)	29.19 (20)	32.63 (23)	32.63 (23)	32.63 (23)
		2	15.52 (7)	15.50 (7)	15.50 (7)	16.49 (7)	16.48 (7)	16.48 (7)
		3	**10.93** (4)	10.37 (4)	10.36 (4)	**11.00** (4)	10.69 (4)	10.69 (4)
		4	/	**7.87** (2)	7.79 (2)	/	8.19 (2)	8.14 (2)
		5	/	7.90 (2)	6.49 (2)	/	**7.42** (2)	6.57 (2)
		6	/	/	**5.69** (1)	/	/	**5.79** (1)
		7	/	/	7.65 (2)	/	/	7.03 (2)
12	**3**	**1**	53.65 (29)	53.64 (29)	53.64 (29)	78.92 (44)	78.92 (44)	78.92 (44)
		2	26.39 (10)	26.25 (10)	26.25 (10)	29.50 (12)	29.44 (12)	29.44 (12)
		3	**20.31** (5)	17.91 (5)	17.89 (5)	**19.95** (6)	19.25 (6)	19.24 (6)
		4	/	**13.68** (3)	13.48 (3)	/	14.33 (4)	14.25 (4)
		5	/	15.93 (3)	10.88 (2)	/	**12.95** (3)	11.47 (3)
		6	/	/	**9.29** (2)	/	/	**9.37** (2)
		7	/	/	15.86 (3)	/	/	11.33 (2)

Table 5: Comparison of optimal *E(W)* and optimal value of *k* for different allocations of n_p versus n_s, in the PAS and SWP scenario.

(iv) For a given warehouse setting (m, E(O), and n_{tot}), moving from a PAS policy to a SWP policy tends to leave the optimal workforce allocation unchanged, with only two exceptions. The differences in $E(W)_{opt}$ between PAS with optimal workforce allocation and SWP with optimal workforce allocation are only slight. This indicates that the achievement of workforce balance in

our setting is mainly influenced by factors that do not depend on the policy used (e.g., picker travel times, setup times).

7. CONCLUSIONS

In this paper we have provided an analytical approach for determining the expected system throughput time for online order batching and sorting situations. Such batching and sorting decisions play a role in many warehouses. Our model builds on earlier results (Le-Duc and De Koster, 2004) for travel and throughput time estimation with online batching, and on Whitt's (1983, 1993) approximate queuing network analysis. The model includes arbitrary distributions for customer order size, picking time, sorting time, and setup times for picking or sorting a batch. It can be used to determine the optimal batch size, for optimal allocation of workforce to the picking and sorting processes, and for a comparison of a SWP versus a PAS operation. The throughput time appears to be a convex function of batch size (expressed in the number of orders). The analysis has shown that the throughput time is minimized if the workforce is allocated to the picking and sorting operations such that the stations are approximately balanced. Comparing SWP with PAS operations depends on the particular situation (i.e., ratio of set-up times to service times, and the reduction in picking time versus the increase in sorting time if we move from SWP to PAS). In our experiments we found that SWP can only marginally improve PAS for a given number of workers, when nearly all personnel is allocated to the picking operation.

This research can be extended in several directions. It is possible to include the effect of different storage strategies or different layouts, as this will only impact the travel time component of the service time in the picking area. Estimations for travel times of a pick batch for class-based storage strategies (ABC storage) have been derived by Le-Duc and De Koster (2007). For other storage strategies no explicit results exist. A second, less straightforward, extension could be to include a zoned picking system with parallel pickers. In such a case a customer order can be picked by multiple pickers simultaneously and an order has to be assembled at the sorting and packing stations.

REFERENCES

Blumenfeld D, 2001. Operations research calculations handbook. CRC Press: Boca Raton, FL.

Caron F., Marchet G., Perego A., 1998. Routing policies and COI-based storage policies in picker-to-part systems. International Journal of Production Research 36, 713-732.

Caron F., Marchet G., Perego A, 2000. Optimal layout in low-level picker-to-part systems. International Journal of Production Research 38, 101-117.

Chew E.P., Tang L.C., 1999. Travel time analysis for general item location assignment in a rectangular warehouse. European Journal of Operational Research 112, 582-597.

Choe K, Sharp GP. Small parts order picking: design and operation. Report TR-89-07, Material Handling Research Center, Georgia Institute of Technology, Atlanta, Georgia; 1991. Available at http://www2.isye.gatech.edu/logisticstutorial/order/article.htm (accessed: July 2006).

Choe K., Sharp G.P., Serfozo R.S., 1993. Aisle-based order pick systems with batching, zoning and sorting. In: R.J. Graves et al. (Eds), Progress in Material Handling Research: 1992. The Material Handling Institute of America: Charlotte, NC, pp.245-276.

De Koster R., Van der Poort E.S., 1998. Routing orderpickers in a warehouse: A comparison between optimal and heuristic solutions. IIE Transactions 30, 469-480.

De Koster R., Van der Poort E.S., Wolters M., 1999. Efficient orderbatching methods in warehouse. International Journal of Production Research 37, 1479-1504.

De Koster R., Le Duc T., Roodbergen K.J., 2007. Design and control of warehouse order picking: a literature review, to appear in European Journal of Operational Research.

Dekker R., De Koster R., Roodbergen K.J., Van Kalleveen H., 2004. Improving Order-Picking Response Time at Ankor's Warehouse. Interfaces 34, 303-313.

Elsayed E.A., Lee M.K., 1996. Order processing in automated storage/retrieval systems with due dates. IIE Transactions 28, 567-577.

Elsayed E.A., Lee M.K., Kim S., Scherer E., 1993. Sequencing and batching procedures for minimizing earliness and tardiness penalty or order retrievals. International Journal of Production Research 31, 727-738.

Gademann A.J.R.N., Van den Berg J.P., Van der Hoff H.H., 2001. An order batching algorithm for wave picking in a parallel-aisle warehouse. IIE Transactions 33, 385-398.

Gademann N., van de Velde S., 2005. Batching to minimize total travel time in a parallel-aisle warehouse. IIE Transactions 37, 63-75.

Gibson D.R., Sharp G.P., 1992. Order batching procedures. European Journal of Operational Research 58, 57-67.

Goetschalckx M., Ratliff D.H., 1988. Order picking in an aisle. IIE Transactions 20, 531-562.

Hausman W.H., Schwarz L.B., Graves S.C., 1976. Optimal storage assignment in automatic warehousing systems. Management Science 22, 629-638.

Heskett J.L., 1964. Putting the cube-per-order index to work in warehouse layout. Transport and Distribution Management 4, 23-30.

Hopp W.J., Spearman M.L., 2000. Factory Physics: Foundations of Manufacturing Management. Irwin/McGraw-Hill, Chicago.

Hwang H., Baek W., Lee M., 1988. Cluster algorithms for order picking in an automated storage and retrieval system. International Journal of Production Research 26, 189-204.

Karmarkar U., 1987. Lot sizes, lead times and in-process inventories. Management Science 33, 409-417.

Karmarkar U., Kekre S., Kekre S., 1985. Lot sizing in multi-item multi-machine job shops. IIE Transactions 17, 290-298.

Karmarkar U., Kekre S., Freeman S., 1985. Lot sizing and lead time performance in a manufacturing cell. Interfaces 15, 1-9.

Le-Duc T., de Koster R., 2005a. Travel distance estimation and storage zone optimization in 2-block class-based storage strategy warehouse. International Journal of Production Research 43, 3561-3581.

Le-Duc T., de Koster M.B.M., 2005b. Determining Number of Zones in a Pick-and-pack Order Picking System. Report ERS-2005-029-LIS, RSM Erasmus University: Rotterdam (The Netherlands).

Le-Duc T., De Koster M.B.M., 2007. Travel time estimation and order batching in a 2-block warehouse, to appear in European Journal of Operational Research.

Marshall K.T., 1968. Some inequalities in queueing. Operations Research 16, 651-665.

Ratliff H.D., Rosenthal A.S., 1983. Orderpicking in a rectangular warehouse: A solvable case of the traveling salesman problem. Operations Research 31, 507-521.

Roodbergen K.J., 2001. Layout and routing methods for warehouses. Doctoral dissertation, RSM Erasmus University: Rotterdam (The Netherlands).

Roodbergen K.J., De Koster R., 2001a. Routing methods for warehouses with multiple cross aisles. International Journal of Production Research 39, 1865-1883.

Roodbergen K.J., De Koster R., 2001b. Routing order-pickers in a warehouse with a middle aisle. European Journal of Operational Research 133, 32-43.

Rosenwein M.B., 1994. An application of cluster analysis to the problem of locating items within a warehouse. IIE Transactions 26, 101-103.

Tompkins J.A., White J.A., Bozer Y.A., Frazelle E.H., Tanchoco J.M.A., 2003. Facilities Planning. John Wiley & Sons, Somerset, NJ.

Whitt W., 1983. The queueing network analyzer. The Bell System Technical Journal 62, 2779-2815.

Whitt W., 1993. Approximations for the GI/G/m queue. Production and Operations Management 2, 114-161.

APPENDIX A: APPROXIMATION FOR THE FIRST AND SECOND MOMENT OF TRAVEL TIME

Approximations for the first and second moment of travel time in a two-block rectangular warehouse and S-shape routing policy have been developed in Le-Duc and De Koster (2007). Their results are valid for both random storage policy and ABC-storage policy without partial-aisle assignment. For a given number of order lines n to be picked on a tour, they show that:

$$E\left[TR_{WA}|n\right] = d(m - \sum_{i=1}^{m}(1-p_i)^n)$$
$$E\left[TR_{CA}|n\right] = 2w_c\left[\frac{m}{2} - \sum_{l=1}^{m/2-1}\left(\sum_{r=1}^{l} p'_r\right)^n\right] \quad \text{(A.1)}$$

where $p'_r = 1-(1-p_{2r-1})(1-p_{2r})$ denotes the probability that pick line r (r=1,…,m/2) is visited. $E\left[TR|n\right]$ can then be approximated by the sum of these two components and two adjustment terms:

$$E\left[TR|n\right] = E\left[TR_{WA}|n\right] + E\left[TR_{CA}|n\right] + E\left[AT_1|n\right] + E\left[AT_2|n\right] \quad \text{(A.2)}$$

The first adjustment term $E\left[AT_1|n\right]$ takes into account the average travel time from the center of the cross aisle to the beginning of the first pick aisle, and the average travel time from the end of the last pick aisle back to the center of the cross aisle. It is given by:

$$E\left[AT_1\right] = 2w_a(1-0.5)^n \quad \text{(A.3)}$$

The second adjustment term $E\left[AT_2|n\right]$ takes into account the correction needed when the last visited aisle in a block is odd (note that, as in Figure 1, the pick aisles are numbered alternately from left to right): in that case, the picker does not traverse the entire aisle but instead makes a U-turn at the last pick position to return to the center of the cross-aisle. The expected value of this correction term is given by:

$$E\left[AT_2|n\right] = 2(0.5)^n \sum_{g\in G:\text{odd}}\left[\Pr(g,n)\left(2d\frac{n}{n+g} - d\right)\right]$$
$$+\left[1-2(0.5)^n\right]\sum_{k=1}^{n-1} 0.5^n \binom{n}{k} \sum_{g\in G:\text{odd}}\left[\Pr(g,k)\left(2d\frac{k}{k+g} - d\right) + \Pr(g,n-k)\left(2d\frac{n-k}{n-k+g} - d\right)\right] \quad \text{(A.4)}$$

The first term represents the expected value of the correction term when the turn occurs in only 1 block, which has a probability $2(0.5)^n$ of occurring. The second term refers to the expected value of the correction term when a turn occurs in both blocks, which has a probability $(1\text{-}2(0.5)^n)$ of occurring. In that case, the probability that k picks fall into one block and (n-k) picks into the other is given by $0.5^n\binom{n}{k}$. The notation *Pr(g,x)* refers to the probability that all x picks fall into exactly g

aisles within a block ($g \in \{G | 1 \le g \le m/2, \text{g is odd}\}$) and is given by $\Pr(g,x) = \binom{m/2}{g}\left(\frac{g}{m/2}\right)^x X(g,x)$, where $X(g,x)$ equals $X(g,x) = 1 - \sum_{i=1}^{g-1}(-1)^{i+1}\binom{g}{g-i}\left(\frac{g-i}{g}\right)^n$ and denotes 1 minus the probability that all x picks fall into less than g aisles, conditional upon the fact that all x picks fall into at most g specific aisles. Using expressions (A.1), (A.3) and (A.4) the first moment of travel time, given a fixed number of order lines n, is obtained from (A.2).

The variance of travel time is approximated by:

$$Var[TR|n] = E[TR^2|n] - \left(E[TR_{CA}|n] + E[TR_{WA}|n]\right)^2 \tag{A.5}$$

The second moment of travel time is approximated by:

$$E[TR^2|n] = d^2 E[J^2|n] + (2w_c)^2 E[L^2|n] + 4w_c d E[JL|n] \tag{A.6}$$

where J refers to the number of aisles visited, and L to the pick line of the farthest visited aisle. The three components of expression (A.6) are given by:

$$\begin{aligned}
E[J^2|n] &= m^2 - \sum_{i=1}^{m}(2m-1)(1-p_i)^n + 2\sum_{i=1}^{m-1}\sum_{r=i+1}^{m}(1-p_i-p_r)^n \\
E[L^2|n] &= (m/2)^2 - \sum_{i=1}^{m/2-1}(2i+1)\left(\sum_{r=1}^{i} p'_r\right)^n \\
E[JL|n] &= \sum_{l=1}^{m/2} l\left[\left(\sum_{r=1}^{2l} p_r\right)^n \left(2l - \sum_{i=1}^{2l}(1-p_i^*)^n\right) - \left(\sum_{r=1}^{2(l-1)} p_r\right)^n \left(2(l-1) - \sum_{i=1}^{2(l-1)}(1-p_i^{**})^n\right)\right]
\end{aligned} \tag{A.7}$$

In this expression, $p'_r = 1 - (1-p_{2r-1})(1-p_{2r})$, $p_i^* = p_i / \sum_{j=1}^{2l} p_j$ and $p_i^{**} = p_i / \sum_{j=1}^{2(l-1)} p_j$.

Using expressions (A.6) and (A.1), $Var[TR|n]$ can be approximated by expression (A.5).

APPENDIX B: APPROXIMATION FOR E(W_q) AT A G/G/1 SERVER (WHITT, 1983)

According to the work of Whitt (1983), the average waiting time of entities in queue at a single-class *G/G*/1 server can be approximated by:

$$E(W_q)_{G/G/1} = \begin{cases} \dfrac{\rho^2(ca^2 + cs^2)}{2\lambda(1-\rho)} & \text{if } ca^2 \ge 1 \\ \dfrac{\rho^2(ca^2 + cs^2)}{2\lambda(1-\rho)} \exp\left\{\dfrac{-2(1-\rho)(1-ca^2)^2}{3\rho(ca^2+cs^2)}\right\} & \text{otherwise} \end{cases} \tag{B.1}$$

In this expression, ca^2 refers to the SCV of interarrival times of entities at the server, whereas cs^2 refers to the SCV of the total service time of entities. If we let Y and S denote the interarrival time of entities and the total service time of entities respectively, we have $ca^2 = Var(Y)/(E(Y))^2$ and $cs^2 = Var(S)/(E(S))^2$. The notation λ stands for the arrival rate of entities, and is given by

$\lambda = 1/E(Y)$. The notation ρ refers to the utilization of the server, and is given by $\rho = E(S)/E(Y) = \lambda/\mu$ where μ refers to the processing rate of the server ($\mu = 1/E(S)$).

APPENDIX C: APPROXIMATION FOR E(W_q) AT A G/G/M SERVER (WHITT, 1993)

Whitt (1993) provides the following approximation for the expected waiting time in general multiserver queues:

$$E(W_q)_{G/G/m} = \phi(\rho, ca^2, cs^2, m)\left(\frac{ca^2 + cs^2}{2}\right)E(W_q)_{M/M/m} \qquad \text{(C.1)}$$

For a multiserver station consisting of *m* servers, the utilization is given by $\rho = \lambda/(m\mu)$ with λ and μ defined as in Appendix B. The exact expression for $E(W_q)_{M/M/m}$ is given by

$$E(W_q)_{M/M/m} = \frac{P(N \geq m)}{\mu\, m(1-\rho)}\ ,\ P(N \geq m) = \left[\frac{(m\rho)^m}{m!(1-\rho)}\right]\zeta \text{ with } \zeta \equiv \left[\frac{(m\rho)^m}{m!(1-\rho)} + \sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\right]^{-1}.$$

$P(N \geq m)$ denotes the probability that the number of customers in the system (*N*) exceeds the number of servers (*m*), and is equivalent to the probability that all servers are busy. The expression for ϕ in (C.1.) is given by:

$$\phi(\rho, ca^2, cs^2, m) = \begin{cases} \left(\frac{4(ca^2 - cs^2)}{4ca^2 - 3cs^2}\right)\phi_1(m,\rho) + \left(\frac{c_s^2}{4ca^2 - 3cs^2}\right)\psi(c^2, m, \rho) & ca^2 \geq cs^2 \\ \left(\frac{cs^2 - ca^2}{2(ca^2 + cs^2)}\right)\phi_3(m,\rho) + \left(\frac{cs^2 + 3ca^2}{2(ca^2 + cs^2)}\right)\psi(c^2, m, \rho) & ca^2 \leq cs^2 \end{cases}$$

with $\gamma(m,\rho) = \min\left\{0.24, \frac{(1-\rho)(m-1)\left[(4+5m)^{0.5} - 2\right]}{16m\rho}\right\}$, $\phi_1(m,\rho) = 1 + \gamma(m,\rho)$,

$\phi_3(m,\rho) = (1 - 4\gamma(m,\rho))e^{-\frac{2(1-\rho)}{3\rho}}$, and

$$\psi(c^2, m, \rho) = \begin{cases} 1 & c^2 > 1 \\ \phi_4(m,\rho)^{2(1-c^2)} & 0 \leq c^2 \leq 1 \end{cases}$$

with

$$c^2 = \frac{ca^2 + cs^2}{2},\ \phi_4(m,\rho) = \min\left\{1, \frac{\phi_1(m,\rho) + \phi_3(m,\rho)}{2}\right\}.$$

APPENDIX D: SCATTER PLOTS FOR VALIDATION EXPERIMENT

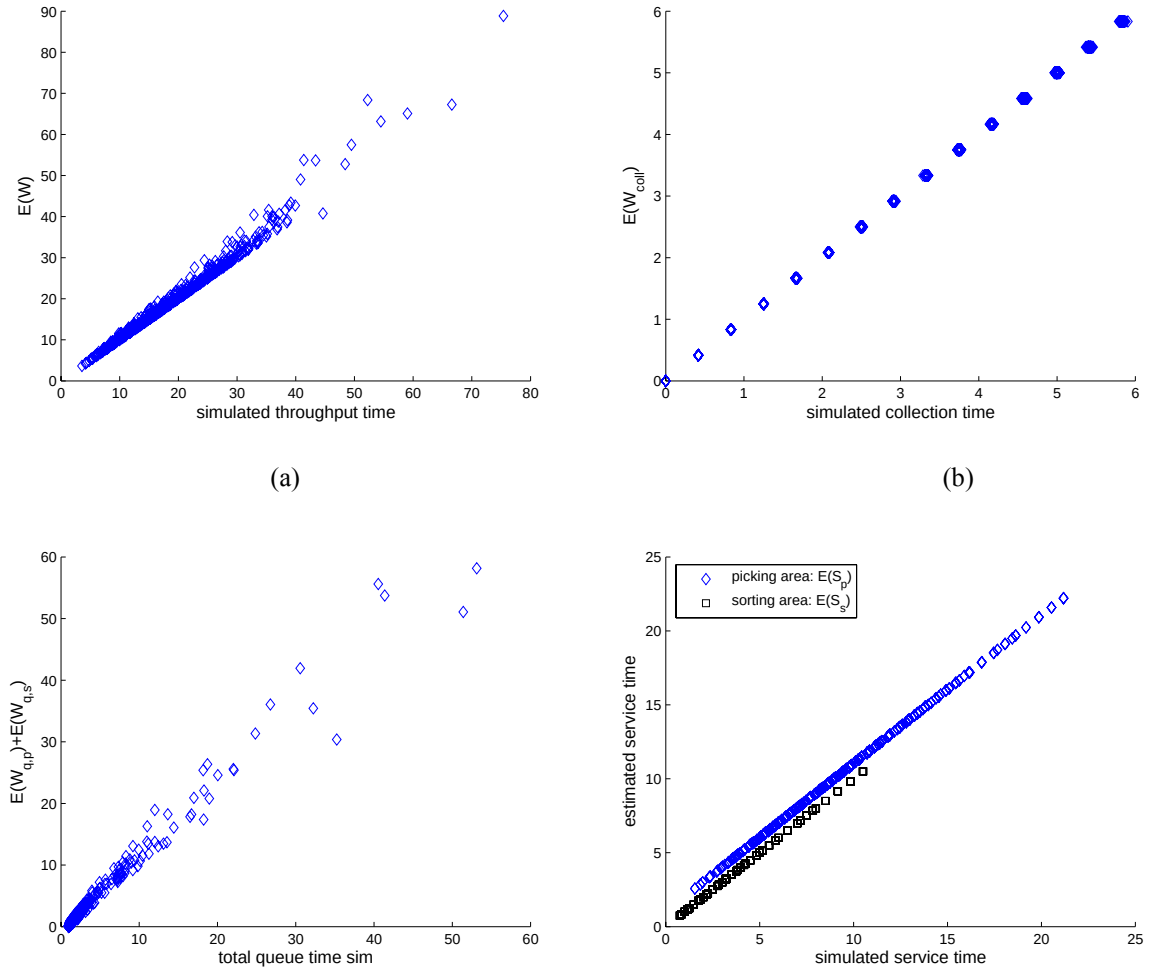


Figure: Scatter plots of the simulation versus the queueing results, in minutes, for (a) the average throughput time, (b) the average collection time, (c) the average total time spent in queue at picking and sorting, and (d) the average service time in picking and sorting area